Decomposition of metabolic network into functional modules based on the global connectivity structure of reaction graph

Hong-Wu Ma1, Xue-Ming Zhao2, Ying-Jin Yuan2 and An-Ping Zeng1,*

1Department of Genome Analysis, GBF—German Research Center for Biotechnology Mascheroder Weg 1, 38124 Braunschweig, Germany and 2Department of Bioengineering, Tianjin University, 300072, Tianjin, People’s Republic of China

Received on August 24, 2003; revised on November 11, 2003; accepted on January 14, 2004
Advance Access publication March 22, 2004

ABSTRACT

Motivation: Metabolic networks are organized in a modular, hierarchical manner. Methods for a rational decomposition of the metabolic network into relatively independent functional subsets are essential to better understand the modularity and organization principle of a large-scale, genome-wide network. Network decomposition is also necessary for functional analysis of metabolism by pathway analysis methods that are often hampered by the problem of combinatorial explosion due to the complexity of metabolic network. Decomposition methods proposed in literature are mainly based on the connection degree of metabolites. To obtain a more reasonable decomposition, the global connectivity structure of metabolic networks should be taken into account.

Results: In this work, we use a reaction graph representation of a metabolic network for the identification of its global connectivity structure and for decomposition. A bow-tie connectivity structure similar to that previously discovered for metabolite graph is found also to exist in the reaction graph. Based on this bow-tie structure, a new decomposition method is proposed, which uses a distance definition derived from the path length between two reactions. An hierarchical classification tree is first constructed from the distance matrix among the reactions in the giant strong component of the bow-tie structure. These reactions are then grouped into different subsets based on the hierarchical tree. Reactions in the IN and OUT subsets of the bow-tie structure are subsequently placed in the corresponding subsets according to a ‘majority rule’. Compared with the decomposition methods proposed in literature, ours is based on combined properties of the global network structure and local reaction connectivity rather than, primarily, on the connection degree of metabolites. The method is applied to decompose the metabolic network of Escherichia coli. Eleven subsets are obtained. More detailed investigations of the subsets show that reactions in the same subset are really functionally related. The rational decomposition of metabolic networks, and subsequent studies of the subsets, make it more amenable to understand the inherent organization and functionality of metabolic networks at the modular level.

Contact: aze@gbf.de
Supplementary Information: http://genome.gbf.de/bioinformatics/

INTRODUCTION

The genome sequencing of a growing number of organisms opens up new horizons for understanding biological processes such as signal transduction and cellular metabolism at the system levels (Hartwell et al., 1999; Ravasz et al., 2002; Kitano, 2002). Because of the general importance of the metabolic network for cellular functions and the fact that it can be faithfully reconstructed from genome information and biochemical literature (Ogata et al., 1999; Overbeek et al., 2000; Karp et al., 2002; Ma and Zeng, 2003a; Forster et al., 2003), it has increasingly become the subject of studies aiming at understanding the organization principles and evolutionary history of biological networks (Jeong et al., 2000; Wagner and Fell, 2001; Podani et al., 2002; Ravasz et al., 2002; Ma and Zeng, 2003b, 2004). The study of genome-based metabolic networks has given remarkable new insights into fundamental aspects of cellular metabolism. One of the important findings is that, like many non-biological complex systems, metabolic networks exhibit typical characteristics of small-world networks, namely a power law connection degree distribution, high cluster coefficients and a short network diameter (Fell and Wagner, 2000; Jeong et al., 2000; Wagner and Fell, 2001). This small-world structure is regarded as one of the design principles of many robust and error-tolerant networks such as the computer network, neural network and certain social and economic networks (Strogatz, 2001; Albert and Barabasi, 2002).
However, the small-world network structure seems to contradict the modular structure of cellular metabolism (Ravasz et al., 2002). In biochemistry, it is well established that modules consisting of several interacting bioreactions or metabolic pathways build discrete functional units of metabolism (Neidhardt et al., 1990; Hartwell et al., 1999). These modules are further nested to form a complex metabolic network. To resolve the apparent contradiction between the small-world structure and modularity organization, Ravasz et al. (2002) proposed a hierarchical modularity model for metabolic networks. According to this model, metabolic networks of organisms are organized as many small but highly connected modules that combine in a hierarchical manner to larger, less cohesive units. Several recent studies using concepts such as the reaction betweenness centrality distribution and the dependency of metabolites, have further verified that metabolic networks are organized in an hierarchical manner (Holme et al., 2003; Gagneur et al., 2003). These results indicate that hierarchic modularity is also an important feature of metabolic networks. Modularity has been shown to be common in the organization of robust and sustainable complex systems (Hartwell et al., 1999). Therefore, identifying the modular organization of metabolic network by certain network decomposition methods can help us better understand the organization principle of complex systems.

Network decomposition is also necessary for functional analysis of metabolic networks. Pathway analysis methods such as elementary flux mode analysis and extreme pathway analysis have been shown to be useful tools for investigating the metabolic capacity and pathway structure of metabolic networks (Schuster et al., 2000; Papin et al., 2002; Stelling et al., 2002; Palsson et al., 2003; Price et al., 2003). These methods are, however, hampered by the combinatorial explosion problem when applied to large-scale networks such as those reconstructed from genomic data. Hence, a decomposition of the network is necessary before performing functional analysis using the methods of pathway analysis mentioned above.

Several methods have been proposed to decompose metabolic networks (Ravasz et al., 2002; Schuster et al., 2002; Holme et al., 2003). Ravasz et al. (2002) use a topological overlap matrix for network decomposition. They classify the metabolites into different subsets. However, from a biological viewpoint, a subset of metabolites cannot sufficiently define a unique functional pathway or module. Reactions catalyzed by enzymes are the functional and regulatory units in a metabolic network. Therefore, a module should be a subset of reactions, and not merely a subset of metabolites. Some metabolites can participate in reactions belonging to different modules. Schuster et al. (2002) proposed a decomposition method by removing the highly connected metabolites. Using this method, they analyzed the metabolic network of a parasitic bacterium Mycoplasma pneumoniae. Compared to other organisms, M. pneumoniae has a relatively small metabolic network that includes less than 200 reactions (Ma and Zeng, 2003a). Using the metabolic network of Escherichia coli as an example, Gagneur et al. (2003) extended the method of Schuster et al. (2002) by serially grouping reactions in order of the degree of the metabolites they connected, thereby finding a hierarchical organization structure. A potential drawback of these degree-based methods is that the global organization structure may be invisible from the local feature such as the connection degree. Considering this drawback, Holme et al. (2003) developed a method to reveal the subnetwork hierarchies of the network by successively removing reactions of high betweenness centrality. With this method, it is generally difficult to obtain subnetworks of similar size, instead, a large subnetwork and many isolated nodes are often found.

In this paper, we propose a new decomposition method based on a global structure analysis of metabolic networks. In structure analysis of metabolic networks in our recent studies (Ma and Zeng, 2003a,b), we removed the connections through currency metabolites to obtain networks that are more suitable for structural analysis. The reaction reversibility information was also considered, because the reaction direction affects the network topology and the biological function of pathways. In this way, a bow-tie global connectivity structure was identified and a giant strong component (GSC) was found that represents the most complex part of the network (Ma and Zeng, 2003b). The bow-tie structure of metabolic networks builds the basis of the new decomposition method. Unlike our previous studies, we use in this work, a reaction graph to represent the metabolic network structure. The metabolic network of E.coli is studied to demonstrate the applicability of this method.

REACTION GRAPH REPRESENTATION AND BOW-TIE STRUCTURE OF THE METABOLIC NETWORK

In order to decompose the metabolic network into reaction subsets, we represent it as a reaction graph in which reactions are defined as nodes, and metabolites as connections between the nodes (Wagner and Fell, 2001). A directed link from reaction node A to reaction node B is added if the product of reaction A is the substrate of B. Currency metabolites in the reaction are removed (for details see Ma and Zeng, 2003a). Thus, the metabolic network is represented as a directed network, e.g. reactions shown in Figure 1(a) can be transformed into a graph represented by Figure 1(b).

We have reconstructed the metabolic networks from genome information for more than 100 fully sequenced organisms (Ma and Zeng, 2003a). Among them, the metabolic network of E.coli is one of the most complex and includes 979 reactions catalyzed by 630 enzymes. Therefore, we use it as an example for network decomposition and for investigating the function of the modules. Before performing the analysis we reduce
the network complexity into two steps: (1) grouping reactions that are only different in currency metabolites such as reaction A + NAD = B + NADH \(_2\) and reaction A + NADP = B + NADPH \(_2\); (2) removing reactions that are only associated with currency metabolites such as ATP + H\(_2\)O = ADP + Phosphate. These reactions can easily be re-added to the corresponding subsets after the decomposition. After reduction, 729 reactions remain in the metabolic network of *E. coli*.

A connectivity analysis showed that the reaction graph also exhibits the ‘bow-tie’ structure (Fig. 2), which was previously found for metabolic networks represented as a metabolite graph (Ma and Zeng, 2003b). For *E. coli*, the GSC includes 194 fully connected reaction nodes (at least one directed path exists between any pair of nodes). The IN subset (62 reaction nodes, corresponding to the substrate subset in the metabolite graph) and OUT subset (250 reaction nodes, corresponding to the product subset in the metabolite graph) contain nodes that can access, or be accessed, by the nodes in GSC. All the other 223 nodes cannot access, or be accessed, by the nodes in GSC; they form a sparsely connected isolated subset (IS). It should be mentioned that 23 relatively small strongly connected components have also been identified, most of which are in the IN and OUT subsets. The remaining strong components are in the IS that contains no more than five reaction nodes.

We compared the metabolite contents of the four subsets from the reaction graph with that directly obtained from the metabolite graph, and found that they are not strictly the same but very similar. As shown previously, the structure of the IS is simple and most of the metabolic conversions occur in subsets GSC, IN and OUT (Ma and Zeng, 2003b). Therefore, we shall consider mainly these three subsets, henceforth, to illustrate the decomposition method proposed.

It should be noted that this bow-tie structure could only be revealed by regarding the metabolic network as a directed network. The clustering coefficient, which was proposed by Watts and Strogatz (1998) and used by Ravasz et al. (2002) to show the modularity of metabolic networks, does not consider the reaction direction and is thus not suitable for directed networks.

**DECOMPOSITION OF THE METABOLIC NETWORK**

For decomposition of the metabolic network, the nodes in the reaction graph should be classified. The first step in classifying nodes in a network is to give a proper distance definition to show the dissimilarity between the nodes. With the distances calculated, we can build an hierarchical tree for finding clusters of functionally closely related reactions. Ravasz et al. (2002) used a topological overlap coefficient in clustering the metabolites. The reaction direction was ignored in this coefficient and only the local connectivity property was considered (only metabolites which are directly connected with the two compared nodes were used in calculating the coefficient). To overcome these drawbacks, we present a new distance definition based on the shortest path length between two reaction nodes. Path length is defined as the number of connections in the shortest path between two reaction nodes. Finding the shortest path requires a complete search of the network, thus the shortest path length is a parameter related to the global
Decomposition of metabolic network

**Fig. 3.** Decomposition of *E. coli* metabolic network. (a) The hierarchical classification tree for reactions in the GSC of the reaction graph: (b) The modular structure of GSC. Different modules are shown in different colors. (c) The modular organization structure of GSC, IN and OUT subsets of the bow-tie structure. (d) Bioreactions (pathways) in subset 5. This figure can be viewed in colour on *Bioinformatics* online.

connectivity structure. The biological basis of this definition is that functionally closely related reactions should also have short path lengths between them. For a directed network, the path length from node A to node B is not necessarily the same as that from B to A. We use the smaller of the two path lengths as the distance between A and B. Because only the nodes in GSC are fully connected and thus have finite path lengths, we first cluster the reactions in GSC. The path length between each pair of reactions in GSC was calculated using Pajek (Batagelj and Mrvar, 1998). A distance matrix for reactions in GSC is then acquired and used to construct the hierarchical classification tree by the neighbor joining method in the Phylip software package (Felsenstein, 1996). The results are shown in Figure 3(a).

Cutting the hierarchical tree at different levels will decompose the network into different subsets. For example, subset 1
in Figure 3(a) can be further subdivided into two or more small subsets, while subsets 3, 4 and 7, 8, 9 can be grouped to form bigger subsets. Because of this hierarchy, there may be multiple optimal decompositions. First, we identify all the branches with ten or more nodes in the tree as subsets, and then arrange the nodes in the same subset so as to bring them close to each other in the reaction graph, as shown in Figure 3(b). Several nodes are near the root of the tree, and thus do not belong to any branch. They remain in the center of the graph forming a core module [the black nodes in Fig. 3(b)]. It can be seen that these nodes are densely connected and, hence, are difficult to be classified into any specific subsets. This initial decomposition can be further improved by: (1) considering the local connectivity to move nodes that are highly connected with the nodes in the core module and other subsets to the core module; (2) investigating the biological function of the reactions, and moving those reactions, which are not closely related to the function of most reactions in the subset, to the core module. The objective of these improvements is to make the obtained subsets densely connected inside and sparsely connected with other parts, and thus make them structurally and functionally independent. This explains why subsets 2 and 5 do not include the whole branch in Figure 3(a).

Reactions in the IN and OUT subsets of the bow-tie structure of metabolic network, are classified using a ‘majority rule’. Reaction nodes that are directly connected to nodes in GSC are placed in the subset to which most of their neighbors in GSC belong; the other nodes are classified into corresponding subsets to which most of their neighbors belong. In this way, all the reactions in the GSC, IN and OUT subsets are separated into different functional subsets as shown in Figure 3(c). As an example, most of the metabolic conversions in module 5 are shown in detail in Figure 3(d). It includes pathways for glutathione synthesis, degradation and usage. The major biological functions of the other subsets were also investigated and are listed in Table 1. Most of the classical pathways defined in textbooks belong to the same subset. For example, in module 2, glutamate, proline, arginine are all synthesized from oxoglutarate, while in module 4, aspartate, lysine, threonine are synthesized from oxaloacetate. However, we also surprisingly find that three classical pathways in the central metabolism, the glycolysis pathway, pentose phosphate pathway and citrate acid cycle, are split into parts in different subsets. This is, however, in consistence with the organization synopsis of E. coli metabolic network proposed by Gagneur et al. (2003). One possible explanation is that the metabolites in these central pathways are used as precursors for the synthesis of different products and are thus placed in different subsets. For example, for reactions in the TCA cycle, the reaction from isocitrate to oxoglutarate is in module 2 because oxoglutarate is the precursor of the glutamate family amino acid; the reactions from malate to isocitrate are in module 4 because oxaloacetate is the precursor for the aspartate family amino acid and aspartate is then used for pyrimidine synthesis. For reactions in the pentose phosphate pathway, all the erythrose 4-phosphate related reactions are in module 8 because it is the one of the precursors for aromatic amino acid synthesis; whereas all the ribose 5-phosphate related reactions are in module 10 because it is the precursor for purine synthesis.

Most subsets include about 20–40 reactions. Networks of this scale can be analyzed easily by pathway analysis methods such as elementary mode or extreme pathway analysis, to better understand their physiological meaning. Subset 10 is the largest with 114 reactions, however, as shown in Figure 3(c), its connectivity structure is not very complex. Therefore, it is not difficult to further decompose it into two or three smaller subsets.

In order to compare our method with previously developed connection degree-based methods (Schuster et al., 2002; Gagneur et al., 2003), we examined the metabolites in the resulting subsets and identified those that are in several subsets.

### Table 1. Reaction subsets from decomposition of genome-based metabolic network of E.coli

<table>
<thead>
<tr>
<th>Subset</th>
<th>Function</th>
<th>Number of reactions</th>
<th>Number of metabolites</th>
<th>Number of internal metabolites</th>
</tr>
</thead>
<tbody>
<tr>
<td>Core</td>
<td>Pyruvate metabolism, glyoxylate metabolism, valine, leucine and isoleucine synthesis</td>
<td>89</td>
<td>134</td>
<td>76</td>
</tr>
<tr>
<td>1</td>
<td>Acetyl-CoA and Succinyl-CoA metabolism</td>
<td>22</td>
<td>53</td>
<td>23</td>
</tr>
<tr>
<td>2</td>
<td>Glutamate and glutamine metabolism, urea cycle, arginine and proline synthesis</td>
<td>38</td>
<td>61</td>
<td>29</td>
</tr>
<tr>
<td>3</td>
<td>Oxaloacetate metabolism and pyrimidine synthesis</td>
<td>47</td>
<td>61</td>
<td>23</td>
</tr>
<tr>
<td>4</td>
<td>Propanoyl-CoA metabolism, threonine, methionine and lysine synthesis</td>
<td>43</td>
<td>79</td>
<td>41</td>
</tr>
<tr>
<td>5</td>
<td>Glutathione metabolism</td>
<td>19</td>
<td>42</td>
<td>20</td>
</tr>
<tr>
<td>6</td>
<td>Glycerate and galactarate metabolism</td>
<td>10</td>
<td>19</td>
<td>8</td>
</tr>
<tr>
<td>7</td>
<td>Glucose, galactose and nucleotide sugar metabolism</td>
<td>37</td>
<td>55</td>
<td>39</td>
</tr>
<tr>
<td>8</td>
<td>Fructose and mannose metabolism, aromatic amino acid synthesis</td>
<td>44</td>
<td>65</td>
<td>35</td>
</tr>
<tr>
<td>9</td>
<td>Glycerone phosphate and glycerolipid metabolism</td>
<td>33</td>
<td>50</td>
<td>26</td>
</tr>
<tr>
<td>10</td>
<td>Pentose phosphate pathway, purine, folate and riboflavin synthesis</td>
<td>114</td>
<td>123</td>
<td>84</td>
</tr>
</tbody>
</table>
These metabolites connect the subsets together and may be regarded as external metabolites, in analogy to the definition of external metabolites for the calculation of elementary modes of the metabolic network (Schuster et al., 2002). Correspondingly, the metabolites that exist in only one subset may be considered as internal metabolites. The total number of metabolites and the number of internal metabolites in every subset are also given in Table 1. We calculated the connection degree (number of reactions a metabolite participated in) for all the metabolites. It was found that most of the highly connected metabolites are also external metabolites. This is in consistence with the results obtained by Schuster et al. (2002) and Gagneur et al. (2003). Actually, most of these metabolites are currency metabolites such as H₂O, ATP, NADH, CO₂, etc. Different from their results, we also found that several internal metabolites such as glucose, galactose and glutathione have a high connection degree (9–10), while several low connection degree metabolites such as homocysteine and glyceraldehyde (degree 3) are external metabolites. From Figure 3(d) it is obvious that if glutathione is removed, the subset will be split into many isolated reactions. This phenomenon has also been observed by Holme et al. (2003) in studying the subnetwork hierarchy by serially removing reactions in order of betweeness centrality. They found that removal of one or more high betweeness centrality reactions often led to a giant component and many isolated reactions, but not several similar size subsets. This is due to the fact that certain highly connected metabolites may be located in the center of periphery subnetworks, in which case, the connection degree based methods may fail to find the proper subsets.

It should be mentioned that the connection degree of metabolites is a local network structure property. The decomposition method presented in this work is based on the bow-tie structure and the shortest reaction path length. Both of them reflect the global connectivity structure of the network. The local connectivity is considered, in our method, in the refinement of the initial decomposition to find proper edges of subsets. Generally, our method first finds subnetworks from the global structure and then adjusts them in detail by investigating the local connectivity. Thus, our decomposition method makes use of both the local and global properties of the network. The resulting subsets are of more adequate size and have inherent biological function. They can serve as a good basis for the modular and functional analysis of metabolic networks.

CONCLUSION

A bow-tie global connectivity structure exists in metabolic networks represented as reaction graphs, as found previously in metabolite graphs. Based on the bow-tie structure of the reaction graph, a new decomposition method is proposed that uses the path length as the dissimilarity measure between two reactions for the construction of a hierarchical classification tree of reactions. This hierarchical tree can be used to classify the reactions into different functionally closely related subsets. Application of the decomposition method to the genome-based metabolic network of E.coli resulted in a rational decomposition of the network into 11 subnetworks (modules) with well-defined biological functions. Compared with methods previously proposed in literature, our method has the advantage of combining both the local and global properties of the metabolic network.

ACKNOWLEDGEMENTS

This work was financially supported through the Braunschweig Bioinformatic Competence Center project ‘Intergenomics’ of the Ministry for Education and Research (BMBF), Germany (Grant No. 031U110A) and by the National Natural Science Foundation of China in an Overseas Young Scientist Cooperation project (No. 20028607) and the key NSFC project (No. 20036010).

REFERENCES


